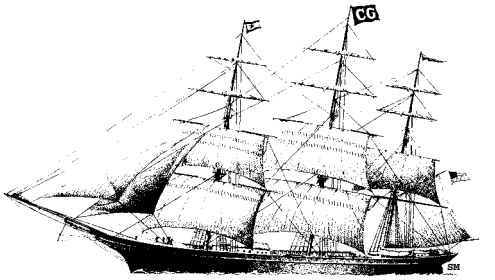
THE CLIPPER GROUP
Explorer™



Navigating Information Technology Horizons℠

Published Since 1996 Report #TCG2004082 October 7, 2004

iWARP Wrings the Overhead Out of Ethernet

Analyst: Anne MacFarland

Management Summary

Like the solitary weightlifter in Olympic competition, enterprise computing was once a self-sufficient endeavor focused on computational power. Benchmarks were predictive of results. Then, business use of computing got more widespread, networking got into the act, cooperation between compute nodes became part of the process, and enterprise computing grew to resemble the manic collaboration of a handball team. Data movement between servers has now become a cornerstone of computing. Marshall McCluhan's pronouncement, *the media is the message,* has been reborn in a less-elegant ***the data transmission is the process*** **and, therein, lies a problem.**

In enterprise computing, as in low-scoring handball (or Red Sox) games, there often seems to be a lot of latency in a process involving devices and networks that theoretically should be able to do more. In IT, the explanation is one of overhead incurred by the processors. Some of this overhead is caused by all the functions that CPUs accomplish besides computation, such as the messaging of distributed applications and the wrapper stacks of TCP/IP, CORBA, or other enabling technologies. However, **most of the CPU overhead comes from the network procedures that are used to safely send data from one server to another – procedures that were set back in the days when such data transfer was not so common.**[1] With the coming of 10-Gigabit Ethernet, the growth of network bandwidth has far outstripped the growth of processing power, as defined by Moore's Law. These new pipes will remain woefully underutilized as servers choke on their own data transmission overhead unless something is done.

The *InfiniBand* (IB) incarnation of Remote Direct Memory Access (RDMA) dealt specifically with the data transfer overhead, introducing an open, standard, but brand new networking protocol. Adoption of IB has been slow and limited. Now, **another incarnation of RDMA, called *iWARP,* has emerged that translates the imperatives of RDMA, as developed in InfiniBand, into Ethernet extensions.** This approach involves merely a swap-out of Network Interface Cards (NICs), and allows reuse of the existing Ethernet networks, skills, management, and optimization. A complete implementation of iWARP can offload 90% of data transmission overhead from the server CPU, reducing latency to near-IB levels and improving system performance.

iWARP is a far more pervasive solution than that provided by limited implementations of InfiniBand or by the specialized protocols of Fibre Channel or proprietary supercomputing connectivity. It allows a vision of a simpler, everything-in-Ethernet network, involving fewer skill sets, a more limited range of infrastructure devices, and a leaner inventory of spare parts. This is attractive to those fed up with IT complexity. Read on for more details.

IN THIS ISSUE

[1] Except, of course, as usual, on mainframes that were designed to run multiple workloads simultaneously, using proprietary internal communication networks such as IBM's *Hypersockets*.

Known Problem, Familiar Path to a Solution

The problem of data transmission latency has been recognized for some time, as can be seen in the *RDMA Chronology* below (in Exhibit 1). The RDMA solution approach has followed a classic pattern of pulling together elements of tested technologies (in this case, elements of server clustering and intermediation/offloading strategies), and involving technology vendors large enough to see the value in *pro bono* work on a common problem for as long as it takes to solve it. The result is a standards-based solution, developed by many experts, completed as InfiniBand, then recompiled (so to speak), and extended, in *iWARP*, for a wider audience on Ethernet.

The character of the solution, an offload of obligations from the operating system and CPU, is also familiar. We have seen this approach in floating point operations and in graphics processing. The co-processor will often be some combination of state engine (for traditional blazing throughput) and programmable silicon (to allow modification over time). Now is the time for this to happen to the process of processor transmission and reception of data.

The Particulars of Data Transfer Process Latency

Let's take a brief look at Ethernet latency, the onerous nature of data transfer between servers, and how the stringent requirements for security and control, as well as pervasive utility, mandate the iWARP set of elements. (See Exhibit 2 on the next page.)

Ethernet traffic between computers has three general sources of latency. One is the TCP/IP stack. Offloading this is a well-developed capability, but, badly done, the offload adds its own latency back in. Well done, it gets rid of only 40% of Ethernet network latency. More latency is caused by the many sources of messaging (for monitoring and management, between devices for coordination) that clutter the network and increase the likelihood of interrupts, which put more work on all the servers using the network.

Data transfer itself is a major source of network latency. The data transfer process produces latency in the following ways:

- The routine *copying* of data from the application into the kernel memory buffer before it is sent to

Exhibit 1 - RDMA Chronology

- 1997 - Intel, Compaq and Microsoft draft the Virtual Interface Architecture (VIA), which becomes one basis for Infiniband (IB).
- 1999 - The Infiniband (IB) Architecture is drafted. No protocol is specified.
- 2001 - The Direct Access File System (DAFS) 1.1 spec. and DAFS API are released.
- 2002 - The Direct Access Transport (DAT) Collective releases user-level and kernel-level Direct Access Programming Libraries, which run today on iWARP as well as IB.
- 2002 - The RDMA Consortium develops specs for iWARP (RDMA over TCP/IP). iWARP 1.0 is released in Oct. 2002 and iWARP Verbs are released in Feb. 2003.
- 2004 - The Interconnect Software Consortium (ICSC) (composed of Fujitsu, HP, IBM, Network Appliance, and Sun) releases an interconnect transport API 1.0 for IB, VIA, and iWARP.

ICSC is working on iWARP version 1.2. NFS-RDMA, the successor DAFS, is moving through the IETF process as is RDMA-over-IP (RDDP).

another server's processor,

- The lack of a way to more persistently assign memory buffer space on the receiving end (known as the sink node), and thus,
- The need for the application to *communicate with resident operating systems* at both ends of the movement path for the appropriate authentication (known as context switching).

The need to copy the payload repeatedly and the need to communicate with the OS kernel can reduce the amount of actual data that gets transmitted down to less than half of what it theoretically should be. They also involve processes that are critical to the control and security of systems. They cannot be blown off to achieve better throughput.

RDMA allows data to be transferred from the memory of a processor in one system directly to the memory of a processor in another system without the need to communicate with the operating system. It registers the communication buffers of both source and sink nodes in intermediary directories via Direct Access Protocol Libraries for user and kernel spaces (uDAPL and kDAPL). You need both libraries to get full latency reduction.

This intermediation, specifically the user level access incarnation, is how server clustering has worked for years, for the techniques that permit transparently fast failover are also good for transparently fast data movement. However, clustering almost always has been platform specific, because the operating system kernels are deeply involved. The amount of work needed to develop an OS-agnostic approach to the problem was what delayed the arrival of InfiniBand. Much of the InfiniBand work has been reused in iWARP, speeding development and facilitating its ratification as a standard.

The full set of iWARP specs (see Exhibit 2, at right) deal with OS disintermediation and the need for persistent I/O buffer awareness on the part of both processors.

Exhibit 2 –
The Six Elements of iWARP

Wire-side elements - these are all header elements

- **RDMAP – Remote Direct Memory Access Protocol** – The basic protocol that defines how the data will be transferred in a Protocol Data Unit (PDU) from the memory of one server to the memory of another server. This protocol is common to both iWARP and Infiniband.
- **DDP – Direct Data Placement Protocol** – the information in that gives the transmission process details, allowing data to be placed directly in an upper-level protocol's receive buffer without intermediate copies.

Note: the attributes below all enable reuse of existing and familiar IT assets.

- **MPA– Marker PDU Alignment** – a framing protocol to adapt DDP to TCP.
- **iSER – iSCSI Extensions for RDMA** – integrates RDMAP with iSCSI. This allows iWARP to handle block-mode traffic.

Software elements – these reside on servers:

- **RDMA verbs**– RDMA verbs enable user-level direct access, allowing the context switch associated with user-to-kernel transition to be avoided. (These are similar to IB verbs, but more are added to leverage the capabilities of Ethernet and to deal with iSCSI).
- **SDP – Sockets Direct Protocol** allows enterprises to deploy existing sockets-based applications over RDMA.

They also deal with TCI/IP and iSCSI protocol overhead. Thus, iWARP has all that is necessary to afford the critical benefits.

iWARP Benefits

The benefits of iWARP are immediate.

- **iWARP offers an open standard approach to server clustering** that will take the need to choose a platform out of the clustering choice, and thus accelerate its use.
- **iWARP is compatible with the message passing interface** (MPI) of highly parallelized supercomputing, and can be used in existing high performance environments.
- **NFS v4, among its many enhancements, is being made RDMA aware**.[2] This will allow it to outperform block-level access at a given wire speed.[3]
- **iWARP makes IP-based storage even more attractive.** ISER allows iSCSI to take advantage of RDMA.
- **FICON can be accommodated** - with bridges.
- Thanks to InfiniBand, **industry-standard RDMA APIs are already available for applications** like *Oracle*, *Windows 2000*, and *Windows 2003*.

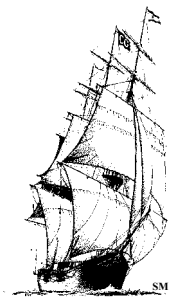
There are also important long-term ramifications.

- **iWARP's low latency meets the high-performance needs** of clusters and grids, the fast-data-access needs of SANs, and the high-throughput needs of NAS with the familiar connectivity enterprises already have on their LANs. The ability to converge networks is a cost savings not to be dismissed lightly, no matter how radical it seems.
- **The disintermediation of the operating system** is consistent with, and fosters the development of scale-out computing, blades, and Linux.
- **Software-as-a-service** heightens the visibility of the network role in enterprise computing. Latency will become a point of competitive disadvantage. The demand for iWARP in the smaller enterprise may come sooner than predicted in traditional adoption patterns.
- Because **the barriers to adoption are relatively low**, adoption should be relatively swift, and volume manufacture rapid. The price reduction curve of iWARP RDMA-NICs (and it's just a swap-out of NICs that is entailed) should follow the precipitous Ethernet model.

As with any efficiency measure based on standards, completeness will be the differentiator between solutions. 10 Gigabit Ethernet is ten times the capacity of Gigabit Ethernet. It is a lot of bigness to think about, and a lot of scale to swallow. It might be tempting to go with a partial solution, but the benefits are so compelling that this would be a mistake. **Fully utilized, iWARP can allow the enterprise to accomplish a lot more with a lot less**. The improvements that it will afford to data services like replication (including backup), data synchronization, and the parallel processing of enterprise search will lead quickly to pervasive use of iWARP in enterprise IT environments. Moreover, for the smaller enterprise that has not adopted fiber channel, iWARP's advantages will be particularly compelling.

Conclusion

Networked computing has become an essential element of the processes of all organizations. It is time to make the networks involved work better. iWARP is a platform-agnostic, OS-independent, low-impact way to do this. It will become a basic tool for all enterprises using 10 Gigabit Ethernet, as *Warp Speed* was for the starship *Enterprise* in *Star Trek*. It may also cause whiplash in those organizations whose competition adopts it first. Think *now* about what your enterprise's iWARP strategy could be, should be, and will be.



[2] NFS-RDMA supports both NFS V3 (in use now) and NFS V4.

[3] NFS clients (the user side) need never enter kernel space to access storage. Because in block storage schemes the file system resides in the kernel, such a bypass is not possible with Fibre Channel.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland **is Director of Enterprise Architectures and Infrastructure Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 28. (Please dial "1-28" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.